International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Classification of Paddy Types using Naïve Bayesian Classifiers

Mie Mie Aung, Su Mon Ko, Win Myat Thuzar, Su Pan Thaw

Information Technology Supporting and Maintenance Department,
University of Computer Studies, Meiktila, Myanmar

How to cite this paper: Mie Mie Aung | Su Mon Ko | Win Myat Thuzar | Su Pan Thaw "Classification of Paddy Types using Naïve Bayesian Classifiers" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.1355-1359, **https://doi.org/10.31142/ijtsrd26585**

IJTSRD26585

ABSTRACT

Classification is a form of data analysis that can be used extract models describing important data classes or to predict future data trends. Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. In classification techniques, Naïve Bayesian Classifier is one of the simplest probabilistic classifiers. This paper is to study the Naïve Bayesian Classifier and to classify class label of paddy type data using Naïve Bayesian Classifier. This paper predicts four class labels and displays the selected impacts attribute of each class label by using Naïve Bayesian classifier. Moreover, this paper can predict the types of paddy for paddy dataset by using other classification methods such as Decision Tree and Artificial Neural Network. Furthermore, this system can be used to predict production rate and display the selected impacts attribute of other crops such as soybeans, corns, cottons. This paper focuses on paddy dataset and decides paddy types are Lasbar or Yar Sabar or Yenat Khan Sabar or Sar Ngan Khan Sabar.

KEYWORDS: *Naïve Bayesian, Paddy types, Classification, Large dataset*

1. INTRODUCTION

Computers are widely used in Education, health, Arts, Humanities, Social Science, Industry, Communication, Government, Administration, Research, Business sectors and Agricultures.

Computer systems are stored large amount of data, are often required not only to retrieve but also to classify that data rapidly in a variety of sequences, combinations and classification. Data mining is one of the computer-aided systems such classification prediction, etc. It refers to extracting or mining knowledge from large amounts of data. Classification is an important technique in data mining. A classification model can also be used to predict the class label of unknown records. The paddy types are identified as definitely Lasbar, definitely Yar Sabar, definitely Yenat Khan Sabar and Sar Ngan Khan Sabar. Paddy types have numbers of instances and numbers of attributes. Paddy data is large dataset. Along with decision trees and neural networks, Bayesian Classifier is one of the most practical and most used learning methods. When to use:

- Moderate or Large Training set available,
- Attributes that describe instances are conditionally independent given classification.

So, this paper is to implement Bayesian Classifier using paddy types.

2. RELATED WORK

Herry Zhang proposed the sufficient and necessary conditions for the optimality of naïve Bayes. He investigated the optimality of naïve Bayes under the Gaussian distribution. W. Zhang and F. Gao described an auxiliary feature method is proposed. It determines features by an existing feature selection method, and selects an auxiliary feature which can reclassify the text space aimed at the chosen features. Then the corresponding conditional probability is adjusted in order to improve classification accuracy. They show that the proposed method indeed improves the performance of naive Bayes classifier. J. Ren proposed a novel naive Bayes classification algorithm for uncertain data. His key solution is to extend the class conditional probability estimation in the Bayes model. Extensive experiments on UCI datasets show that the accuracy of naive Bayes model can be improved by taking into account the uncertainty information. Toon Calders and Sicco Verwer investigated how to modify the naive Bayes classifier in order to perform classification that is restricted to be independent with respect to a given sensitive attribute.

3. THEORETICAL BACKGROUND

Data mining is the process of digging or gathering information from various databases. The data mining should have been more appropriately names knowledge mining from data. Data mining involves the use of sophisticated data analysis tool to discover previously unknown, valid patterns and relationships in large datasets. These tools can include statistical models, mathematical algorithms, and machine learning methods.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameter to examine the data. They include association, sequence or path analysis, classification, clustering and forecasting.

3.1 KINDS OF DATA MINING

A number of different data stores on which mining can be performed. In principle, data mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World Wide Web. Advanced database systems include object-oriented and object-relational databases, and specific application-oriented databases, such as spatial databases, time-series databases, text databases and multimedia databases. The challenges and techniques of mining may differ for each of the repository systems.

3.2 TYPES OF DATA MINING

In general, data mining tasks can be classified into two categories: descriptive and predictive.

1. Descriptive tasks: The objective is to derive patterns that summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require post-processing techniques to validate and explain the results.
2. Predictive tasks: The objective of these tasks is to predict the value of a particular attribute based on the other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for the explanatory or independent variables.

Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. The kinds of data mining are Association Analysis, Classification and prediction, Cluster Analysis, Outlier Analysis, and Evolution Analysis. Data mining systems can be categorized according to various criteria, as follows.

- Classification according to the Kinds of Database Mined
- Classification according to the Kinds of Knowledge Mined
- Classification according to the Kinds of Techniques Utilized
- Classification according to the Application Adapted

4. CLASSIFICATION

Classification is a data mining (machine learning) technique used to protect group membership for data instances. Data classification is a two-steps process. In the first step, a model is built describing a predetermined set a data classes or concepts.

Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes. The model is constructed by analyzing data tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. In the context of classification, data tuples are also referred to as samples, examples or objects.

In the second step, the model is used for classification. First, the predictive accuracy of the model or classifier estimated. If the accuracy of the classifier is considered acceptable, the model can be used to classify future data tuples for which the class label is not known. Such data are also referred to as "unknown" or "previously unseen" data.

Any classification method uses a set of features or parameters to characterize each object, where these features should be relevant to the task at hand. We consider here methods for supervised classification, meaning that a human expert both has determined into what classes an object may be categorized and also has provided a set of sample objects with known classes. This set of known objects is called the training set because it is used by the classification programs to learn how to classify objects.

There are two phases to constructing a classifier. In the training phase, the training set is used to decide how the parameters ought to be weighted and combined in order to separate the various classes of objects. In the application phased, the weights determined in the training set are applied to a set of objects that do not have known classes in order to determine what their classes are likely to be.

If a problem has only a few (two or three) important parameters, then classification is usually an essay problem. For example, with two parameters one can often simply make a scatter-plot of the feature values and can determine graphically how to divide the plane into homogeneous regions where the objects are of the same classes. The classification problem becomes very hard, though, when there are many parameters to consider.

Not only is the resulting high-dimensional space difficult to visualize, but there are so many different combinations of parameters that techniques based on exhaustive searches of the parameter space rapidly become computationally infeasible. Practical methods for classification always involve a heuristic approach intended to find a "good-enough" solution to the optimization problem.

A classification model can also be used to predict the class label of unknown records. A classification model can be treated as black box that automatically assigns a class label when presented with the attribute set of unknown record. The classifier design can be performed with labeled or unlabeled data. Using a supervised learning method the computer is given a set of objects with known classification and is asked to classify an unknown object in the information acquired by it during the training phase.

The classifier design can be performed with labeled or unlabeled data. Using a supervised learning method the computer is given a set of objects with known classification and is asked to classify an unknown object based on the information acquired by it during the training phase.

4.1 METHODS INCLUDES IN CLASSIFICATION

Classification methods are needed for processing the huge quantities of data generated by modern astronomical instruments. Some of methods include in classification are:

- Decision trees
- Bayesian classification
- Classification by back-propagation
- Classification based on concepts from association rule mining

4.1.1 BAYESIAN CLASSIFICATION

Bayesian Classification is based on Bayes' Theorem. Bayesian Classifiers are useful in predicting the probability that a sample belongs to a particular class or grouping. This

technique tends to be highly accurate and fast, making it useful on large databases. Depending on the precise nature of the probability model, Naïve Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naïve Bayes models uses the method of maximum likelihood; in other words, one can work with the Naïve Bayes model without believing in Bayesian probability or using any Bayesian methods.

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Naïve Bayes classifiers often work much better in many complex real-work situations than one might expect.

An advantage of the Naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

4.1.2 NAÏVE BAYESIAN CLASSIFIERS CHARACTERISTIC

Naïve Bayesian Classifiers generally have the following characteristics.

They are roust to isolated noise points because such points because such points are averaged out when estimating conditional probabilities from data. Naïve Bayes Classifiers can also handle missing values by ignoring the example during model building and classification.

They are robust to irrelevant attributes. If X_i is an irrelevant attribute, then P (X_i/Y) becomes almost uniformly distributed. The class conditional probability for X_i has no impact on the overall computation of the posterior probability and C) Correlated attributes can degrade the performance of Naïve classifiers because the conditional independence assumption no longer holds for such attributes.

4.1.3 NAÏVE BAYESIAN CLASSIFICATION EQUATION

P(X) is constant for all classes, so finding the most likely class amounts to maximizing $P(X/C_i)$ $P(C_i)$. P (C_i) is the prior probability of class i. If the probabilities are not known, equal probabilities can be assumed. Assuming attributes are conditionally independent:

$$P(X_k/C_i) = \Pi^n{}_{k=1}\, P(X_k/C_i)$$

P (X_k/C_i) is the probability density function for attribute k. P (X_k/C_i) is estimated from the training samples. Estimate P (X_k/C_i) as percentage of samples of class i with value X_k. Training involves counting percentage of occurrence of each possible value for each class. Also use statistics of the sample data to estimate $P(X_k/C_i)$. Actual form of density function is generally not known, so, Gaussian density is often assumed. Training involves computation of mean and variance for each attribute for each class Gaussian distribution for numeric attributes:

$$P(X_k/C_i) = \frac{1}{\sqrt{2\pi\sigma_{ci}}} e^{\frac{(x_{k-\mu c_i})^2}{2\sigma_{c_i}}}$$

Where,
μC_i is the mean of attribute k observed in samples of class C_i.
δC_i is the standard deviation k observed in samples of class C_i.

5. IMPLEMENTATION

In this training data set, there are 13 attributes. All of the attributes are normal attributes. There are 4 classes, Lasabar, Yasabar, Sar Ngan Khan Sabar and Yenat Khan Sabar. The following table describes name of attributes and description of these attributes.

Table1. Attribute Information

	Attribute Name	Description
1.	Plant genus	Etmahta, Lethyaysin, Ngasein, Midone
2.	Plant Lifetime	120-125,125-130,140-145, 130-135,135-140, 115-120
3.	Plant height	4.0'-4.5',3.0'-3.5',3.5'-4.0',4.5'-5.0',2.5'-3.0', 5.0'-5.5', 5.5'-6.0'
4.	Plant spike	10-20,6-8,9-10,7-9,9-11,8-10,5-7,2-3, 4-6,10-15
5.	Spike length	9.5",11",12",10.5",11",9", 8.5",10.8",12",10"
6.	Seed number in a spike	240,117,160,95,155,130, 140,150,170,234,250
7.	Seed weight(1000) (g)	25.5,21,26,29,28,25,22, 24.5,25.5,19,16,23
8.	Productive rice(%)	45,40,50,55,45,60,54,53, 41,64,51,37, 90,49,60
9.	Amilo (%)	23,25,19,18,20,21,30,26,24
10.	Rice quality	Kyilin, Baikphyupar, Nauk
11.	Consumption	Fair,Soft,Good,Hard
12.	Light response	Yes or No
13.	Production rate	80-100,40-70,60-70,60-80, 100,100-150,40-60, 30-50,100-120

5.1 ALGORITHM OF PADDY TYPES BASED ON NAÏVE BAYESIAN CLASSIFICATION

Algorithm: Naïve Bayesian Classification. Predict class membership probabilities, such as the probability that a given sample belongs to a particular class.

Input: Database, C, of the selected training samples dataset and unknown data X.

Output: Predict class membership, Lasabar or Yasabar or Sar Ngan Khan Sabar or Yenat Khan Sabar.

Method:
K=total record count of training sample dataset C:
for(i=0;i<Ck-1;i++)
{
if(C.record (i).cell (Result).value==Lasbar)
LasabarCount++;
else if (C.record(i).cell(Result).value==Yasabar)
YasabarCount++;
else if (C.record(i).cell(Result).value==Sar Ngan Khan Sabar)
Sar Ngan Khan Sabar Count + + ;else if (C.record(i).cell(Result).value==Yenat Ngan Khan Sabar)
Yenat Ngan Khan SabarCount++;
}
totalLasabarProb= LasabarCount/k;
totalYasabarProb=YasabarCount/k;
totalSar Ngan Khan SabarProb=Sar Ngan Khan SabarCount/k;
totalYenat Khan SabarProb=Yenat Khan SabarCount/k;
m=total record count of testing sample data except ID and Result fields;
n=total cells count in each record of testing sample dataset T expect ID and Result fields;
for(i=0;i<Tm-1;i++)
{
LasabarProb=1;
YasabarProb=1;
Sar Ngan Khan SabarProb=1;
Yenat Khan SabarProb=1;
for(j=0;i<Tn-1;i++)
{
LasabarCount=getCount (j,T.record(i).Cell(j).value, Lasabar);
LasabarProb *=LasabarCount/ LasabarCount;
YasabarCount=getCount (j,T.record(i).Cell(j).value, Yasabar);
YasabarProb *=YasabarCount/ YasabarCount;
Sar Ngan Khan SabarCount=getCount (j,T.record(i).Cell(j).value, Sar Ngan Khan Sabar);
Sar Ngan Khan SabarProb*= Sar Ngan Khan SabarCount/ Sar Ngan Khan SabarCount;
Yenat Khan SabarCount=getCount (j,T.record(i).Cell(j).value,Yenat Khan Sabar);
Yenat Khan SabarProb*= Yenat Khan SabarCount/ Yenat Khan SabarCount;
}
if(LasabarProb> totalLasabarProb)
display result as Lasabar;
else if (YasabarProb> totalLasabarProb)
display result as Yasabar;
else if (Sar Ngan Khan SabarProb > totalSar Ngan Khan SabarProb)
display result as Sar Ngan Khan Sabar;
else if (Yenat Khan SabarProb > totalYenat Khan SabarProb)
display result as Yenat Khan SabarProb;
}
Procedure getCount (colIndex, colVal, result)
k=total record count of training sample dataset C;
count=0;
for(i=0;i<Ck-1;i++)
{
If(C.record(i).cell(colIndex).value==colVal&&C.record(i).cell(i).cell(Result).value==result)
count++;
}
return count;

5.2 SYSTEM FLOW DIAGRAM

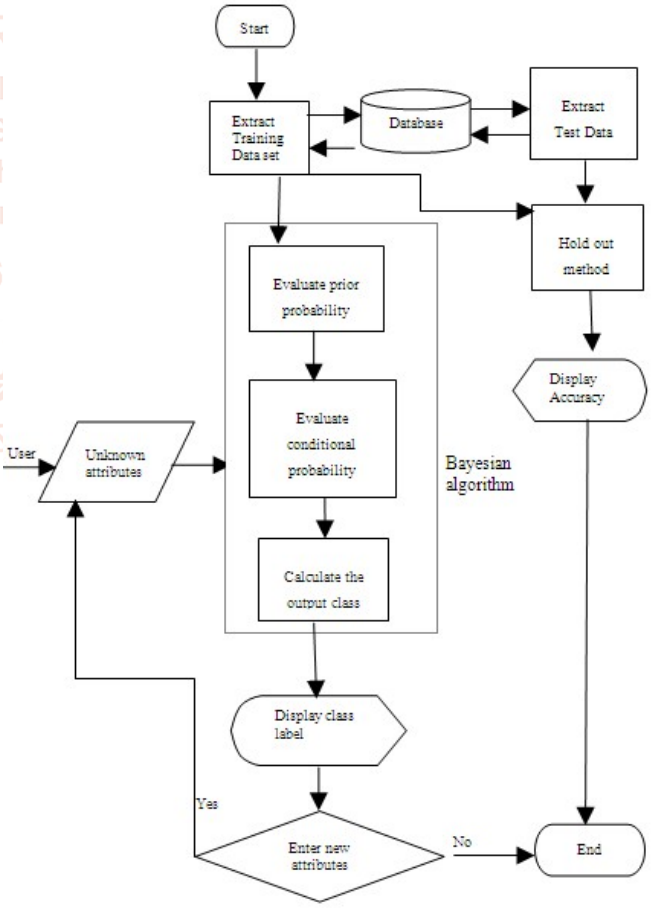

Figure1. Overview of System

5. CONCLUSION AND FURTHER EXTENSION

This paper has presented generating of classification from large datasets. This approach demonstrates efficiency and effectiveness in dealing with many datasets for classification. And then considered the classification problem by using Naïve Bayesian Classification. The accuracy of dataset can

also assessed using Hold-out method. The relative performance of the Naïve Bayesian Classifier can serve as an estimate of the conditional independence of attributes. This paper will extend Naïve Bayesian classifier to work on the other data sets. Moreover, it can circulate the paddy type dataset by using other classifiers such as Decision Tree and Artificial Neural Network.

REFERENCES

[1] Caruana, R. and Niculescu-Mizil, A.: "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning, 2006. (available online PDF (http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.122.5901& rep=rep1&type=pdf))

[2] George H. John and Pat Langley (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345. Morgan Kaufmann, San Mateo.

[3] Harry Zhang "The Optimality of Naive Bayes". FLAIRS2004 conference. (available online: PDF (http://www.cs. unb.ca/profs/hzhang/publications/FLAIRS04ZhangH. pdf))

[4] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques" Morgan Kaufmann, 2001.

[5] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Elsevier 2006, ISBN 1558609016. This part of the lecture notes is derived from chapter 6.4 of this book.[6] Yong Wang, Julia Hodges, Bo Tang "Classification of Web Documents Using a Naive Bayes Method", Mississippi State, MS 39762-9637, 2003.

[6] J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung, ¯Naive Bayes Classification of Uncertain Data,. no. 60703110.

[7] M. Kantardzic, Data Mining - Concepts, Models, Methods, and Algorithms, IEEE Press, Wiley-Interscience, 2003, ISBN 0-471- 22852-4.

[8] Pu Wang, Jian Hu, Hua-Jun Zeng, Lijun Chen, and Zheng Chen, "Improving Text Classification by Using Encyclopedia Knowledge," ICDM '07 Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, 2007.

[9] Toon Calders, SiccoVerwer, ¯Three naive Bayes approaches for discrimination-free classification., Data Min Knowl Disk, 2010.

[10] W. Zhang and F. Gao, ¯Procedia Engineering An Improvement to Naive Bayes for Text Classification,. vol. 15, pp. 2160–2164, 2011.